Validation of Automatic Vehicle Location Data in Public Transport Systems

Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 318 (2015) 31–51

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Validation of Automatic Vehicle Location Data in Public Transport Systems

Stephen Gilmore and Daniël Reijsbergen

Laboratory for Foundations of Computer Science
University of Edinburgh
Edinburgh, Scotland

Abstract

Performance metrics for public transport systems can be calculated from automatic vehicle location (AVL) data but data collection subsystems can introduce errors into the data which would invalidate these calculations, giving rise to misleading conclusions. In this paper we present a range of methods for visualising and validating AVL data before performance metrics are computed. We illustrate our presentation with the specific example of the Lothian Buses Airlink bus, a frequent service connecting Edinburgh city centre and Edinburgh airport. Performance metrics for frequent services are based on *headways*, the separation in space and time between subsequent buses serving a route. This paper provides a practical experience report of working with genuine vehicle location data and illustrates where care and attention is needed in cleaning data before results are computed from the data which could incorrectly reflect the true level of service provided.

Keywords: Public transport measurement and modelling, data cleaning, headway computation

1 Introduction

Modern engineered systems are reflexive. Through instrumentation and sensors, they collect data on their function and performance which is used to assess their progress and safe operation. Transport systems work in this way: a modern bus fleet has richly-instrumented vehicles which report their latitude and longitude, speed and heading. This data is streamed back over a data connection to an automatic vehicle location tracking system which feeds other systems such as real-time arrival prediction for bus passengers.

Judging by recent advances in the field of adaptive systems, it would seem that the future offers us a vision of self-organising, self-healing systems regulated and kept in check by their data-collection subsystems. Unfortunately, these data-collection subsystems are themselves often complex systems, with their own faults and problems, and intrinsic limitations to their engineering. It is not until one starts working with such subsystems that some of these problems begin to become evident. These problems increase in significance when regulators begin to calculate

http://dx.doi.org/10.1016/j.entcs.2015.10.018

performance metrics for public transport services from historical Automatic Vehicle Location (AVL) data traces.

Determining service performance has until recently been done by human observers in place by the side of the road recording vehicle departures and applying intelligence and experience to interpret and record events. This approach has the benefit of ensuring that data is scrutinised before performance measures are calculated. In contrast, in the context where human intelligence is not applied (as in automated processing of historical AVL data traces), errors of interpretation can occur, and it is these errors which are our concern here.

In this paper, we present an experience report on the use of AVL data for obtaining headway and frequency measurements. The AVL data is provided to us by the Lothian Buses company, based in Scotland and operating an extensive bus network in Edinburgh. We consider the specific example here of the Airlink bus service, connecting Edinburgh city centre and Edinburgh airport. An undesirable feature of a frequent service is *clumping*, where two or more buses remain close to each other for an extended period. For this reason *headway*, the separation between successive buses, is an important metric for regulations and service operators.

In particular, we discuss the computation of headway measurements to evaluate the performance of bus routes in terms of specific measures of punctuality. We use a range of methods to visually represent both the data and the computed headways, including a visualisation tool that uses the Google Maps API and which was developed at the University of Edinburgh [1].

The AVL data which is made available to us records the position of each bus in the fleet in terms of Ordnance Survey of Great Britain eastings and northings measurements, which can be easily converted to more familiar latitude and longitude coordinates. The AVL data is specific to a particular bus, as determined by a unique bus identifier called a *fleet number*. The assignment of buses to routes is captured in a schedule which is drawn up before the bus service begins for the day, but may change without notice during the day in response to operational problems. This uncertainty about which buses are in service and which are not gives rise to part of the problems of interpreting the AVL data before metrics are computed.

The remainder of this paper is structured as follows. We first discuss the visualisation of AVL data in Section 2, before moving on to the isolation and removal of data errors in Section 3. We discuss the visualisation of headway data in Section 4 and the use of headway measurements in service level agreements in Section 5. We discuss related work that uses the same data or tools in Section 7, and conclude the paper in Section 8.

2 The value of data visualisation

We are undertaking a modelling exercise which is strongly rooted in data. One vitally important sub-task here is to learn about the data, its scope, and its limitations. In our work with the Lothian buses data we have developed a visualisation tool which allows us to literally view the data in geographical context, against a

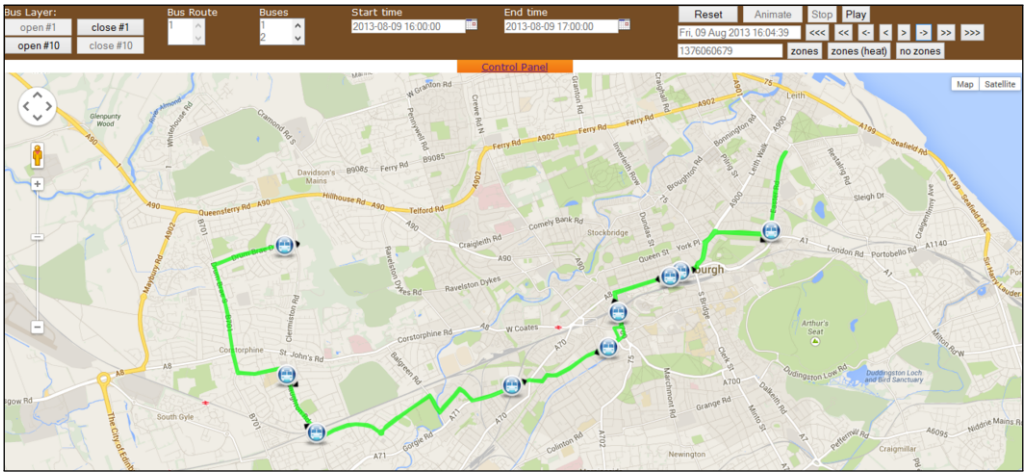

Fig. 1. The user interface allows the user to select bus routes of interest and dates and times of interest and step through the data to see events which occurred in the selected part of the city of Edinburgh.

map of the city of Edinburgh.

This visualisation tool allows us to revisit historical trace data on bus movement and to play or single-step through the data, visualising only those bus services which are of interest. In this way, it brings the data to life, making it easy to confirm that one is looking at the right bus routes. It is shown in Figures 1, 2 and 3.

The visualisation tool has no predictive power, it can only render measurement data. Neither has it any logical, inferential, deductive or verification capacity. Nonetheless, it was very valuable in allowing us to find some significant errors in the data, which we then set about removing in a systematic process of data cleaning.

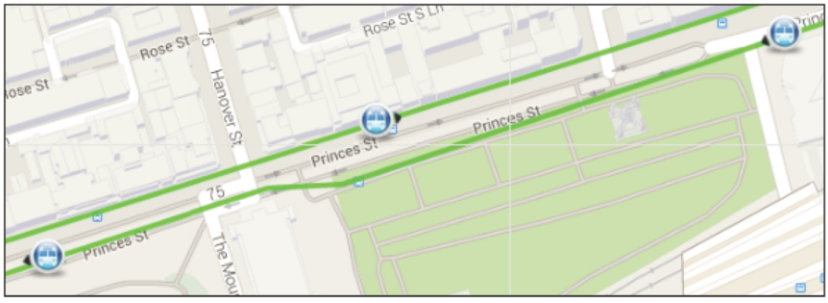

Fig. 2. The data can be accurate enough to confirm the direction of the bus by inspecting visually the side of the road on which it is driving.

In working with data on bus movement from Lothian Buses we are fortunate to have useful domain knowledge about what buses can and cannot do. For example, we know that buses cannot teleport, so when we see that some Edinburgh buses appear to visit Wales (as in Figure 4) we know that this is only a phantom GPS result from the data-collection subsystem which we can discard. Similarly, Edinburgh buses are not amphibious, so measurement data which has them swimming about in the Firth of Forth is also to be discarded. Finally, Edinburgh buses cannot fly, so when we see data which when rendered on the map seems to show them flying

over the rooftops like Ron Weasley's magical car, we know not to believe this. Our visualisation helps us to make sense of this kind of erroneous data by showing that it is a straight line between the final stop on a vehicle's last journey of the night and the bus depot where they are housed overnight.

These erroneous position reports come from vehicles which are not in service, or from measurement sensors which have not been powered down as completely as they should have been, or they are artefacts caused by interpolation in the system trying to fill in data points to compensate for the gap in the data caused by the location-tracking subsystem being switched off at the end of the day's use for a vehicle. However, the data does not record which buses are in service and which are not, so if using the data for purposes other than those for which it was being collected – as we are here – then we need to interpret with care and attention and clean the data to remove erroneous measurements such as these before calculating any measures of interest.

2.1 Visualisation of single bus trajectories

The validation of the service provision which we will conduct depends fully on the data, its quality and completeness, and our interpretation of the data. In order to

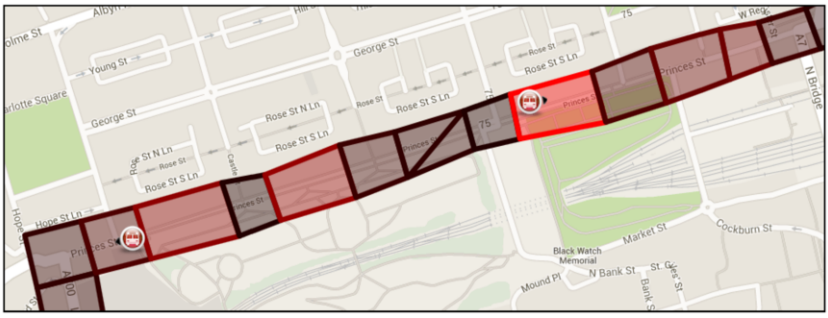

Fig. 3. A heat map representation of Edinburgh city centre showing the patches along Princes Street where buses have the longest sojourn time.

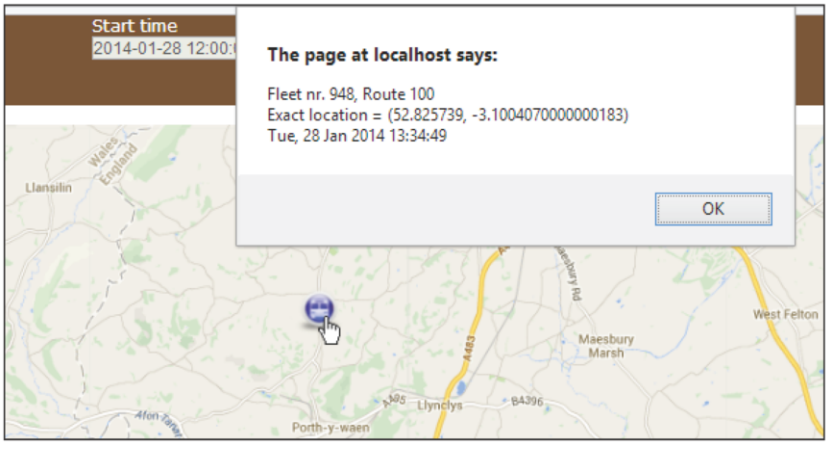

Fig. 4. The visualisation tool can be used to identify buses that appear in a peculiar place, such as a field near the Anglo-Welsh border.

provide ourselves with as thorough an understanding of the data as possible, we developed different views on the data, each of which had value in establishing some understanding of the data and bringing us insights which we found useful.

Our first visualisation, shown in Figure 5, rendered the data in a conventional map view. This was useful in helping us to see that this bus was providing the Airlink service on the day of interest, but it did not use the time content of the measurement trace.

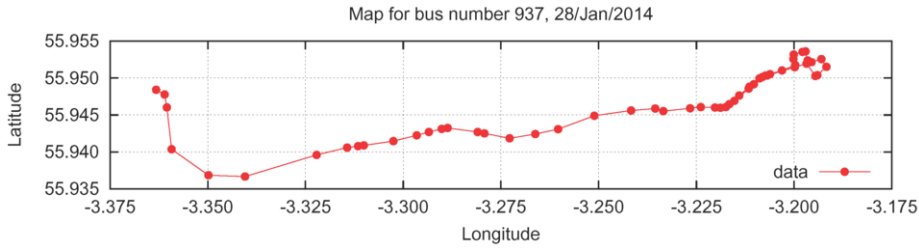

Fig. 5. Sample AVL data from one bus from the airport to the city centre, rendered in the conventional map view which plots latitude against longitude. This view abstracts from both the timing of events and their relative ordering in time.

Our second visualisation, shown in Figure 6, represented time in the abstract sense of *subsequency* in that we used different colours to represent phases of the journey which happen successively. Colours are assigned in a fixed order beginning with red and continuing with orange, yellow, green, blue, indigo and violet before returning to red again. From this visualisation we can see that the bus is travelling from the airport in the west to the city centre in the east, but not at what time of the day (or night) this journey occurred.

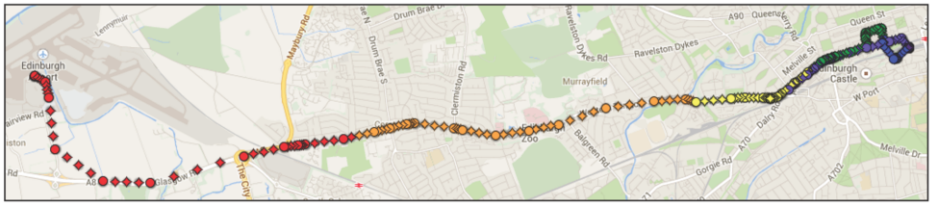

Fig. 6. A visualisation of the airport bus route showing buses on the route from Edinburgh airport in the west to Edinburgh city centre in the east. This visualisation represents sample AVL data from one bus from the airport to the city centre. The colouring imposed on points allows us to determine the direction of the journey.

To allow us to see the journey more clearly it is sometimes helpful to fill in the route in a little more detail by interpolating between the data points. Figure 7 presents such an interpolation. Depending on the use to which this interpolation is put, the measurement errors which are introduced by "cutting corners" as we see in Figure 7 might or might not be problematic.

Finally, Figure 8 shows the AVL data as a time series. In this view, latitude and longitude are plotted separately against time. This has the advantage of allowing us to see when the bus journey happened and to identify positions where it is stationary for long periods of time, which was much more difficult to see in the map view. Against this, our intuitions about where in the city the bus is at any

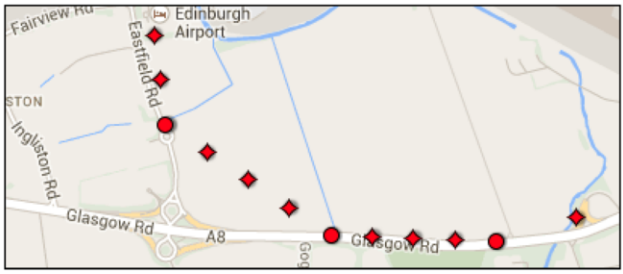

Fig. 7. Some simple interpolation is computed, but naively. The placement of these interpolation points suggests that the bus made the transition from Eastfield Road to Glasgow Road by driving across a field instead of navigating the roundabouts and joining via the slip road, as it of course did.

point in time are lost, because we have moved away from the map view. Thus, this view is complementary to the others which we have seen and provides us with a supplementary understanding of the data, rather than replacing the views which we have seen.

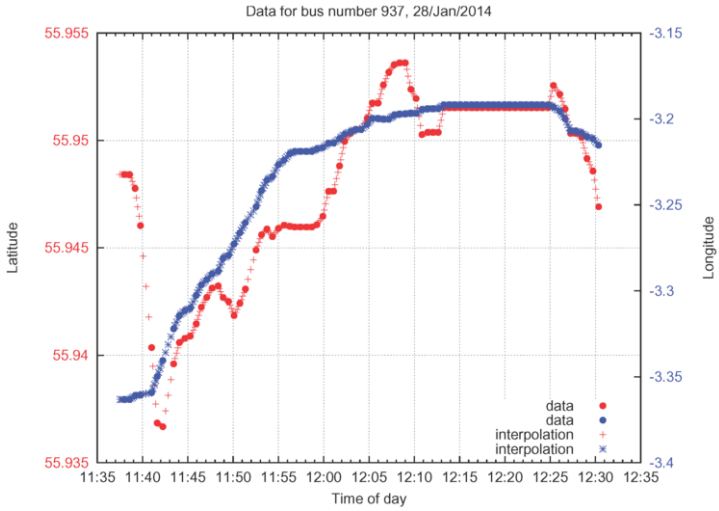

Fig. 8. Sample AVL data from one bus from the airport to the city centre, rendered as a time series. The bus is stationary when neither latitude nor longitude are changing as in this graph between 12:13 and 12:25. The stationary point at 55°57′05.5″N, 3°11′30.3″W is the Airlink bus stop on Waverley Bridge in Edinburgh city centre.

Different views on the data have given us different insights but the identification of collective behaviour remains elusive. In a scenario where events depend not on the behaviour of individuals, but on the behaviour over the long run, or an aggregate measure obtained from a collection of observations, then representing a single individual trace is of little interest. More profound insights come from aggregating individual behaviours to look for trends and patterns.

2.2 Visualisation for collective systems

If we wished to learn the topology of the Airlink bus route in order to identify how and where it turns in order to execute the return journey then this collective view is much more helpful than the individual views which we have seen previously. If

we simply plot all observations of a bus location which we are given, as we do in Figure 9, then this maps out the route without interpolation or approximation (up to the resolution of the GPS data available).

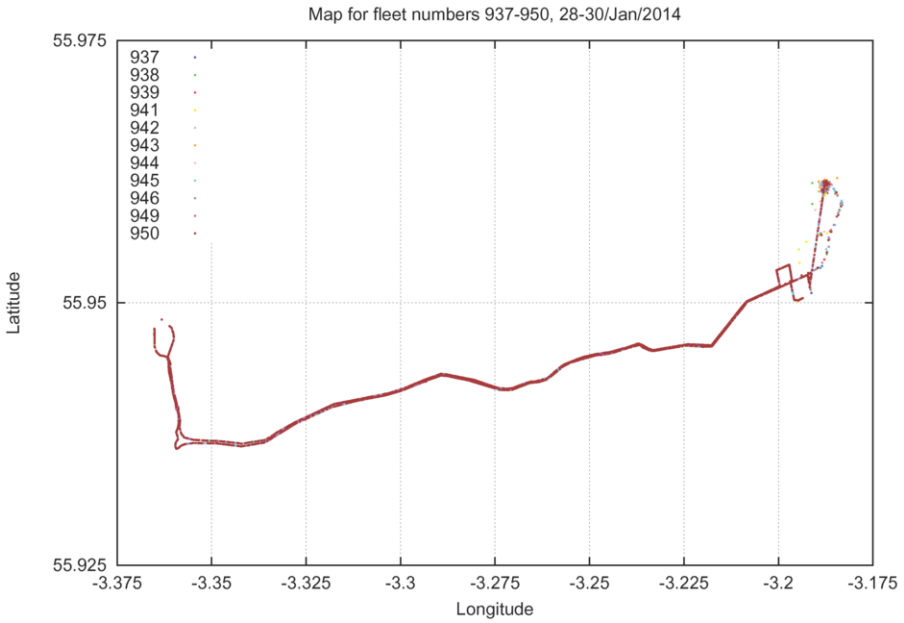

Fig. 9. Latitude and longitude data from eleven buses for two days

Deviations from the planned route can be seen in this view. In this view it is possible to determine that some roads are travelled relatively infrequently (in this case, once per day, the journey from the bus depot to the start of the route, and once per day the journey from the end of the route to the bus depot at night). Occasional diversions from the planned route can occur due to difficult-to-predict events such as traffic accidents, vehicular breakdowns, or even instances of extreme weather conditions. These deviations would also show up in this view, provided that the deviation from the route is long enough that the position of the bus is recorded during the deviation.

Another long-run collective view of the data would be a *heat map*, allowing us to identify where in the city buses spend most of their time (as detected by noting more observations in this area than in others). To achieve this, we place a regular grid over the map of the city with a counter for each square in the grid. We increment the counter every time we see a GPS measurement placing a bus in this square, up to a ceiling of 100 observations per square. Mapping these numbers to a colour spectrum, we see that more-frequently-occupied squares will show up as being hotter than the less-frequently-occupied squares. This might confirm (or refute) our expectations about where delays occur along the route. Figure 10 shows such a view for our data.

From this we can see that the least-frequently travelled part of the route is the journey to the depot in Annandale Street (in the top right-hand corner) because

there are very few observations of buses in this region: only one or two observations for each bus over a two day period. We can also infer that the faster part of the route is on the Glasgow Road leading to the airport (in the bottom left-hand corner). There are more observations here than for the depot, and there is no possible branching off the route here so fewer observations in this region must come from the buses travelling faster here.

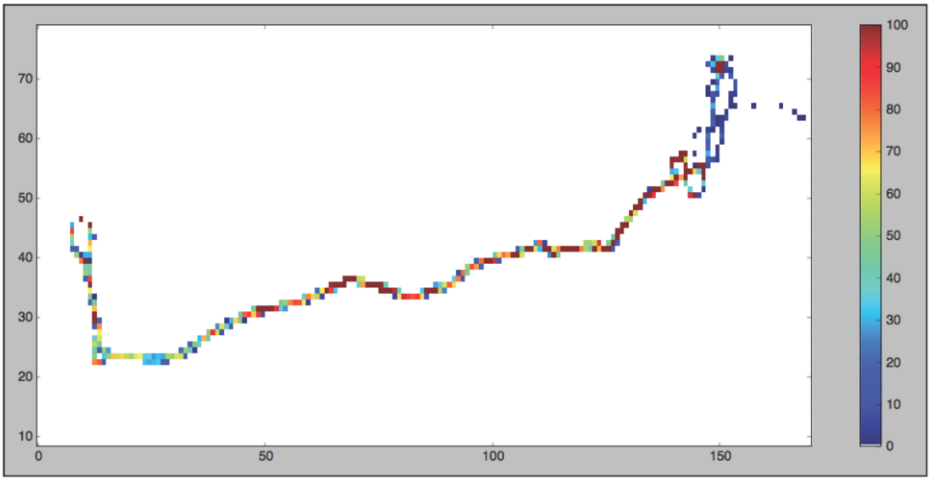

Fig. 10. Heatmap data from eleven buses for two days

3 Isolating errors in data

GPS data can contain both errors of omission and errors of inclusion. Figure 11 demonstrates both of these. The GPS data misses observations on Market Street because this street falls under the GPS shadow of tall buildings on the Mound, a steep hill climbing upwards from Princes Street. This is seen at the bottom of Figure 11 near the centre where there are no data points on Market Street until the roundabout with Cockburn Street and Waverley Bridge.

Figure 11 also contains spurious data points which appear to have been generated by interpolation of observations between Waverley Bridge and the Lothian Buses garage on Annandale Street (note that these are reported data points given to us by the bus company, not like the interpolated data points which we introduced in Figure 7). These manifest themselves as a straight line on the map with interpolated points cutting across York Place and East London Street with no apparent regard for road layout.

The timestamps associated with these data points are either all from the early morning (04:30) when the service starts or last thing at night (23:55) when the service ends. Because of this we believe that these data points are an artefact of cold starts or powering down of the GPS tracking hardware.

Once identified and isolated, erroneous GPS data can be removed using a GPS track editor such as GPSprune [2], as shown in Figure 12. The application shows

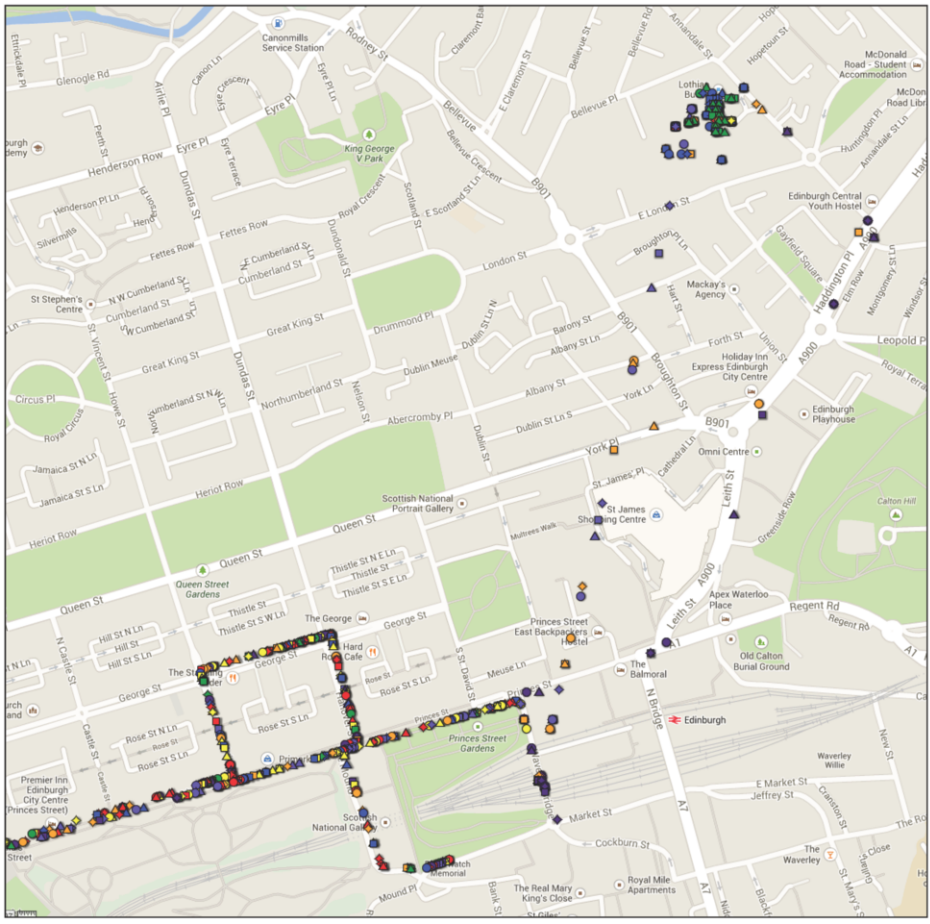

Fig. 11. GPS data for the Airlink bus showing Princes Street, Waverley Station and the Lothian Buses garage in Annandale Street.

derived statistics on the GPS track as well as showing the track in context in a map view, and relating positions on the route to their height.

Using this tool we can conveniently eliminate the erroneous early-morning and late-night interpolated data points. This is a manual editing process, but it is made much more convenient because we can see the data points in context in a standard map view. We can define a region geometrically and then eliminate all of the points which fall within that region. Once this process is complete we are left with a clean data set where all of the erroneous data which we could identify has been eliminated, allowing us to progress on to considering the performance measure of interest (headway). As before, we use visualisation to help us to gain greater insights into the data.

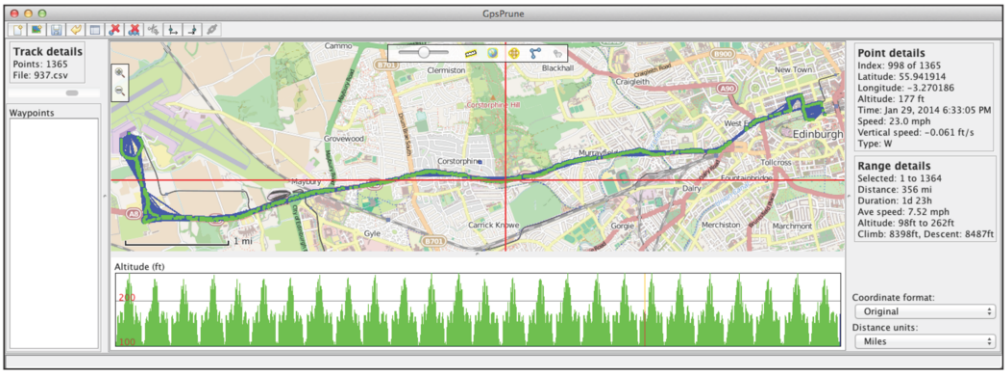

Fig. 12. GPS data being edited in the GPSprune application.

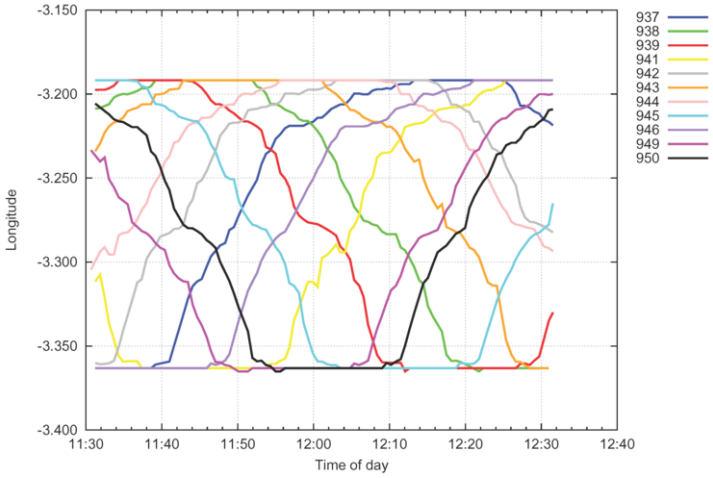

Fig. 13. Average longitude data from eleven buses for one hour

4 Visualising headway

To allow us to approach the computation of headway-based metrics we can begin to look at the collection of eleven buses serving the Airlink bus route. Focusing in on a period of one hour between 11:30 and 12:30, and considering only longitude data as a proxy for progress along the route (because the journey from the city centre to the airport is mostly roughly east-to-west) we can obtain from Figure 13 a sense of headway as separation in time between successive buses.

Looking at the same eleven buses serving the Airlink route over a different granularity of two days, a different pattern emerges. We begin to obtain a sense from Figure 14 of days of service for this collection of buses in the fleet punctuated by overnight absences from service when the buses are stored in the garage.

Note that it is not obvious from published timetable information that the bus operation should follow a day-night pattern. The Airlink bus service runs 24 hours a day and in principle any of the buses from the fleet could be used at any time of

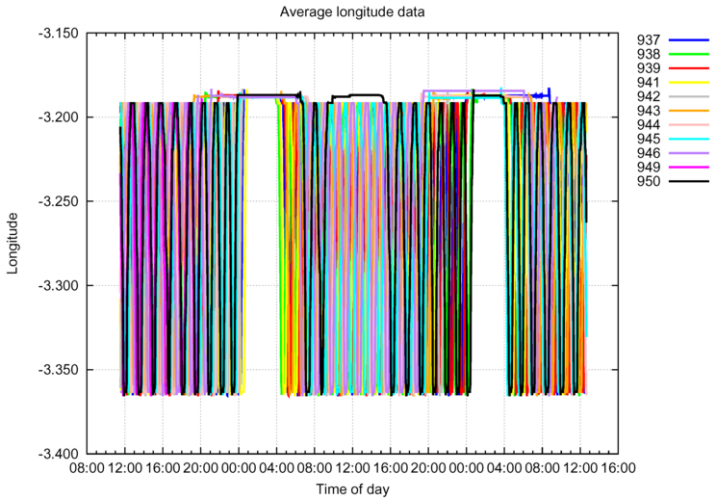

Fig. 14. Average longitude data from eleven buses for two days

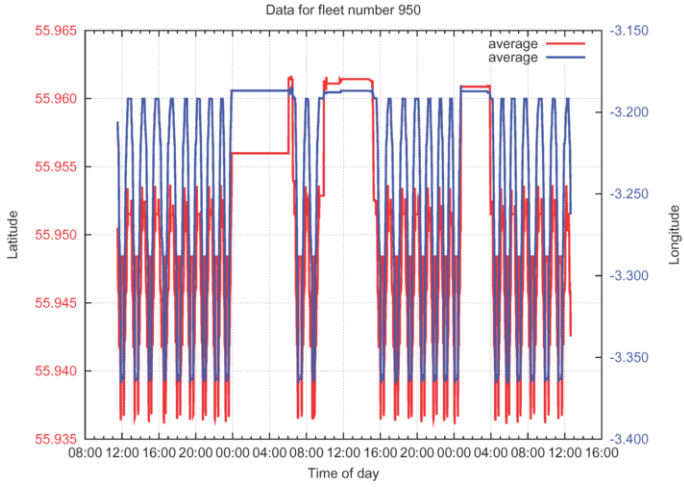

Fig. 15. Average latitude and longitude data from bus number 950 in the fleet over two days. During its absence from service in the middle of the second day this bus is in the Lothian Buses garage on Annandale Street.

day. In practice, one set of buses is used for the daytime service and another set is used for the night-time service.

Understanding whether or not a bus is in service is another important aspect of data cleaning when computing headways. Being widely-separated from a bus which is not in service is much less important than being widely-separated from a bus which is in service.

It is only when we appreciate the day-night pattern that we can notice buses which are not following the pattern. Isolating the data for bus number 950 in the fleet in Figure 15 we can see that it does not follow the established pattern because its second day of service is punctuated by an absence (from approximately 09:30 to 16:00) where it was not serving the Airlink bus route. (We can discover separately

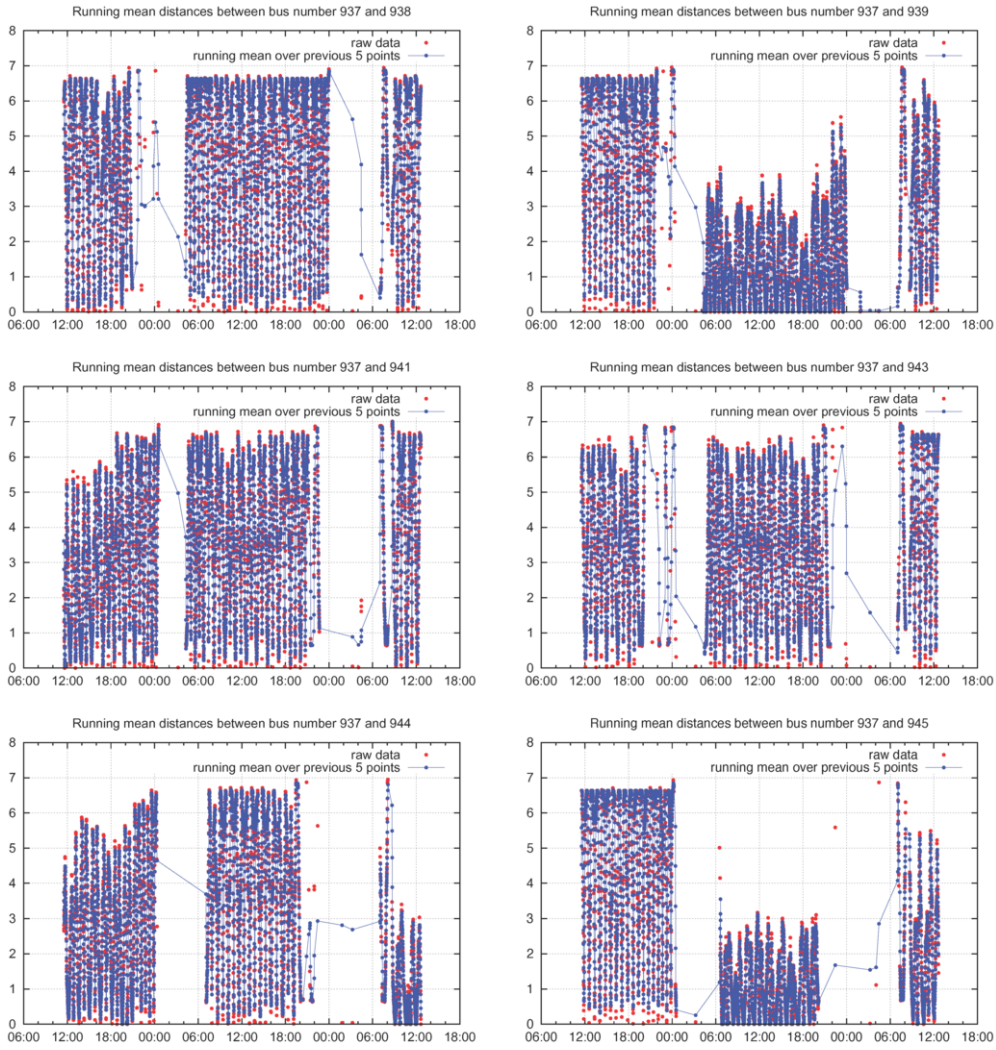

Fig. 16. Time series running mean distances in miles between fleet number 937 and fleet numbers 938, 939, 941, 943, 944, and 945.

that the bus was taken to the garage for an unknown reason such as some kind of mechanical repair to the vehicle, or perhaps a routine service.)

4.1 Spatial separation of service instances

As an alternative to considering headway as separation in time, we could consider headway as separation on route, or at least separation in GPS position. We have used the Haversine function to calculate the spatial distance of one bus from another, as determined by their latitude and longitude, giving their great-circle distances, as commonly used in navigation and spherical trigonometry.

In Figure 16 we present the Haversine distances in miles from fleet number 937,

as a function of time across three days of observation data. This has the benefit of providing a succinct summary of relative bus movement. We can see for each of the three days of data presented that bus number 937 is not close to one of the other buses for an extended period. In each day it moves close to and away from other buses as they make their journeys to and from the airport. The bus which remains closest to fleet number 937 is fleet number 945 which is rarely more than three miles away from fleet number 937 on our second day of observation. Nonetheless, even in this case we can see that the buses are not 'clumping' as the relative distance between the two buses oscillates between zero and three miles throughout the day.

Although it has the benefit of being relatively easy to compute, the Haversine distance is not the ideal distance measure to use to truly calculate the distance between buses. Firstly, it does not take road layout and route into account, and will consistently under-estimate the distance between buses because it calculates the distance "as the crow flies". Secondly, it does not take direction of travel into account: two buses near to each other but on opposite sides of the road, will be reported as being close whereas semantically, they may be several miles apart. Finally, because of variations in speed limits over the route, the same distance denotes a different temporal separation at different parts of the route.

A more relevant distance metric would then be *temporal displacement*, asking "How many minutes has it been since the previous bus was where this bus is now?" This temporal separation is more important to passengers of the system, because regular temporal separation between bus arrivals reduces a passenger's risk of waiting at a bus stop for an unexpectedly long time. Unfortunately these more significant semantic distance measurements which take into account road layout, route, direction of travel, and temporal separation are more difficult to compute than the simple Haversine distance.

The overall result which we would hope to see for a well-running service is spatio-temporal separation where buses are not often close to each other for an extended period, according to some relevant semantic definition of closeness. Such a metric has some merits. It allows for reasonable adaptation to problems in service delivery (so that, for example, buses can on some occasions be close to others for an extended period, as can happen). However, it is not one of the metrics which has been defined by the regulators in this instance.

5 Headway-based service-level agreements

We now consider the service-level agreements which have been identified by the regulators for the service, as published by the Scottish Government. Firstly, there are two classes of bus service identified by regulators: *frequent* and *non-frequent*. Frequent buses depart at least every 10 minutes. Our concern in this paper is only with frequent services. The Airlink bus service is a frequent service departing at least every ten minutes between 04:00 and midnight.

A key characteristic of frequent services is that regulators are not primarily interested in timetable adherence, but rather in the amounts of time between bus

arrivals — the *headways*. We focus on two of the three punctuality metrics for frequent services identified in the guidance document on Bus Punctuality Improvement Partnerships by the Scottish Government [3]; all three are related to headways.

(i) *Six or more buses* will depart from the starting point within any period of 60 minutes on 95% of occasions.

(ii) The interval between consecutive buses departing from the starting point will *not exceed 15 minutes* on 95% of occasions.

The first of these is a requirement on the *frequency* of departures, the second specifies the maximum allowable *headway* between buses.

These service-level agreements are themselves statements about *collective* behaviour. They do not specify that particular individual instances of the service must be correct, but that, viewed as a collection of observations, a large percentage of this collection (in this case, 95% of it) must be satisfactory according to the regulations.

Punctuality is important for regulators but it is of great value to passengers too. The importance of punctuality is such that it has been observed that the negative impact on passenger satisfaction of a decrease in punctuality can outweigh the positive effects of increasing the number of departures per day [4].

5.1 Determining satisfaction of service-level agreements

Through visualisation we have been able to explore various aspects of the available data and investigate problems with the data which need to be resolved but the most important collective system metrics of frequency and headway have not yet been fully explored.

A modelling tool such as Traviando [5] allows us to process trace data and to compute measures of interest over the trace. The primary purpose of Traviando is to act as a post-mortem simulation trace debugger, diagnosing problems with simulation models through statistical, structural, invariant-based and model-checking analysis of output traces. However, because Traviando works with timed trace output, it is possible to invoke it on measurement data such as our time series of GPS observations of bus positions, even before a simulation model is constructed. Figure 17 shows headway observations which have been obtained in this way.

The linear regression across this time series is centred on 476.13 secs, which is approximately 8 minutes, and comfortably less than the 15 minute interval between consecutive buses which is required by the regulator. Furthermore, we observe in Figure 17 that headways of over 15 minutes (900 seconds) are rarely observed, in this case only once in 90 observations.

We define a finite-state process to convert the departure data into a form where we can compute the *frequency* requirement that at least six buses should depart every hour. The process represents a forgetful observer, who counts departures, but forgets departures which happened more than one hour ago, as in Figure 18.

That is, the observer notes the occurrence of each departure of a bus from the

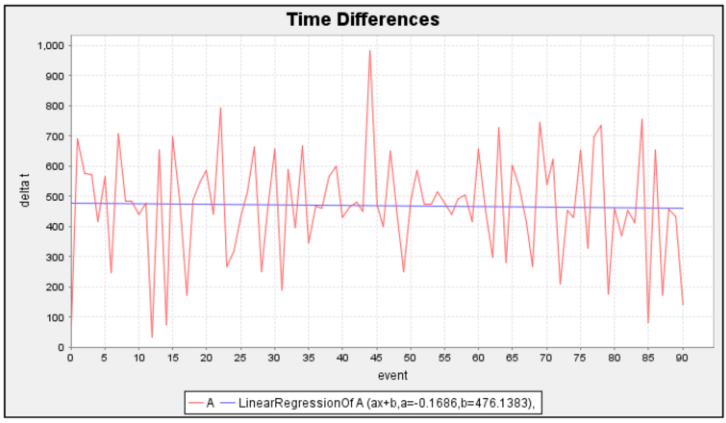

Fig. 17. Headway observations plotted using Traviando as time differences

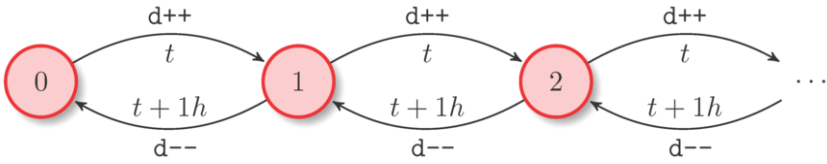

Fig. 18. Observers counting departures should note the times of departures and forget departures which are more than one hour old (at $t + 1h$).

start of the journey and records the time that this event occurred. An hour after any observed departure the departure event is discounted as being outside the relevant window as defined by the Traffic Commissioner regulations. This – not altogether straightforward – process is a reactive system which changes state in response to two types of events: bus departures and clock expiration.

This process allows us to track the frequency metric relating to bus departures, as seen in Figure 19 plotted as a counter value-event trace using Traviando. During the day the observations lie between 6 (the minimum allowable value) and 10. The low period on the second day corresponds to the time when bus number 950 in the fleet was in the garage. Time is abstracted away in this view although the relative ordering of events is maintained. This has determined that regulation (i) above has been satisfied across this observation period.

6 Simulation model

The forgetful observer automaton in Figure 18 provides us with a conceptual model of the process which we use to record (and forget) departures in order to compute the six-buses-per-hour metric. However, in order to analyse the system more deeply we need to develop a simulation algorithm and generate traces to estimate the probability distribution of numbers of departures of buses per hour along the route. The simulation algorithm is the same for each departure point on the route, only

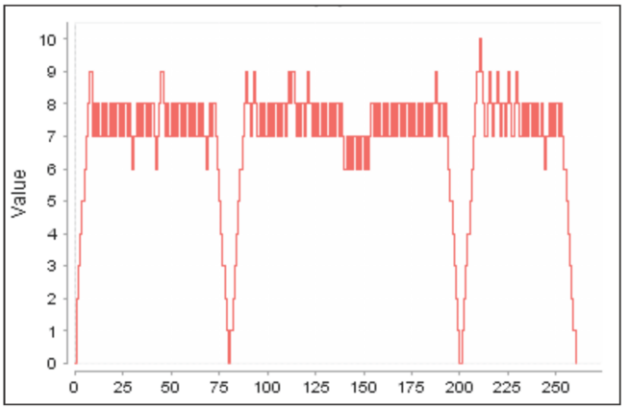

Fig. 19. The frequency metric of buses in the past hour, as obtained from the data for the George Street bus stop.

the numerical parameters of the algorithm are changed to differentiate between the bus stops.

Our goal is to simulate the Z-process, recording (and forgetting) departure events. The simulation algorithm is presented in Algorithm 1, the *runZ* simulation algorithm. The goal of the algorithm is to produce a simulation trace X of (t, z) observations of the time variable t and the variable counting departures per hour, z. In addition, the algorithm computes the probability distribution π_k for integer values of k, giving the long-run probability of there being k departures per hour.

The *runZ* simulation algorithm simulates events forwards in time. Departure events can occur over the course of an eight-hour day. After this time no other departures will occur and the only changes to the z counter will be decrements reflecting departures which happened one hour ago.

The algorithm is guided by an event list, given by the variables L_0 and L_1. L_0 is the time of the next departure event. L_1 is the time when the oldest recorded departure should be forgotten. The simulation algorithm has two cases depending on the relative values of these times.

- If $L_0 < L_1$ then we increment the z counter and record that this departure should be forgotten in one hour's time ($t + 3600$ seconds) either directly as L_1 or in a set of recorded departures D. We then choose the time of the next departure to give a new value for L_0 if we have not already reached the eight-hour limit for departures.
- If instead $L_1 < L_0$ then we decrement the z counter and choose L_1 as being the time of the earliest recorded departure event in D.

Figure 20 presents ten sample runs (X) of the *runZ* simulation algorithm at the Airport, Zoo and George Street bus stops on the journey from the airport to the city centre (at George Street). After the initial transient warm-up period in the first hour the process lies in the range $\{7, 8, 9\}$ at the Airport stop, in the range $\{6, 7, 8, 9\}$

Algorithm 1. The *runZ* simulation algorithm

Require: μ, σ_ϵ, θ

1: $\forall k \in \mathbb{N} : \tau_k := 0;$
2: $D := \emptyset$; $X := \{(0,0)\}$;
3: $u :=$ `random`(); ▷ `random`() draws a number uniformly from $[0,1]$
4: $\epsilon := \sigma_\epsilon \cdot \Phi^{-1}(u)$; ▷ Φ is the Gaussian cumulative distribution function
5: $\epsilon_{\text{prev}} := 0$;
6: $z := 0$; $t := 0$; $t^* := 0$;
7: $L_0 = \max\{0, \mu + \epsilon\}$; $L_1 = \infty$;
8: **while** $L_0 < \infty \ \vee \ L_1 < \infty$ **do**
9: **if** $L_0 < L_1$ **then** ▷ The next event is a departure
10: **if** $t > 3600 \ \wedge \ L_0 < \infty$ **then**
11: $\tau_z := \tau_z + L_0 \quad t$;
12: $t^* := t^* + L_0 \quad t$;
13: **end if**
14: $z := z + 1$; $t := L_0$; $\epsilon_{\text{prev}} := \epsilon$; ▷ Increment z to record the departure
15: $\epsilon := \sigma_\epsilon \cdot \Phi^{-1}($`random`$())$;
16: **if** $t < 8 \cdot 3600$ **then** ▷ Departures stop after 8 hours
17: $L_0 := \max\{0, t + \epsilon + \theta \cdot \epsilon_{\text{prev}} + \mu\}$; ▷ Choose the next departure
18: **else**
19: $L_0 := \infty$; ▷ No more departures take place
20: **end if**
21: **if** $L_1 = \infty$ **then**
22: $L_1 := t + 3600$; ▷ Set the next departure to forget
23: **else**
24: $D := D \cup \{t + 3600\}$; ▷ Add this to the set of departures to forget
25: **end if**
26: **else** ▷ The next event is forgetting a departure
27: **if** $t > 3600 \ \wedge \ L_0 < \infty$ **then**
28: $\tau_z := \tau_z + L_1 \quad t$;
29: $t^* := t^* + L_1 \quad t$;
30: **end if**
31: $z := z \quad 1$; $t := L_1$; ▷ Decrement z to forget the departure
32: **if** $D = \emptyset$ **then**
33: $L_1 := \infty$; ▷ No departures to forget
34: **else**
35: $L_1 = \min D$; $D := D \backslash \{L_1\}$; ▷ Choose the first departure to forget
36: **end if**
37: **end if**
38: $X := X \cup \{(t, z)\}$; ▷ Record the value of z at time t
39: **end while**
40: $\forall k \in \mathbb{N} : \pi_k := \frac{\tau_k}{t^*}$;
41: **return** $(\pi_k)_{k \in \mathbb{N}}$, X

at the Zoo stop, and in the range $\{6, 7, 8, 9, 10\}$ at the George Street stop. Clearly, the variance of the process increases as the bus travels along its route from the Airport to the Zoo to George Street.

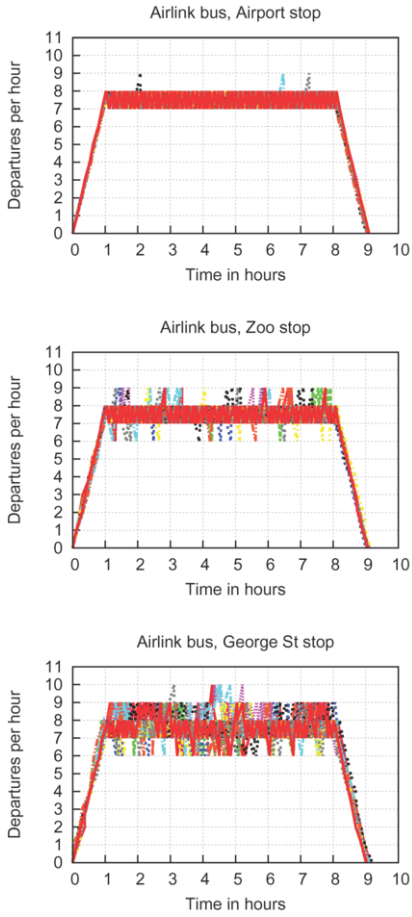

Fig. 20. Ten sample runs of departures per hour for stops on the Airlink route.

Figure 21 presents the probability distributions π_k of the Z process at the Airport, Zoo and George Street bus stops. As we would expect, these distributions reflect the behaviour that the variance of the service increases along the route, as has been typically observed by others in other contexts for other bus routes. However, the probability of the problematic case of fewer than six departures per hour is negligible at all of the bus stops along the route.

7 Related work

The presented data analysis and visualisation methods have been used in several recent papers. In [6], the model checking tool Traviando was used to perform correctness checks on bus journey time data obtained by scraping the Edinburgh

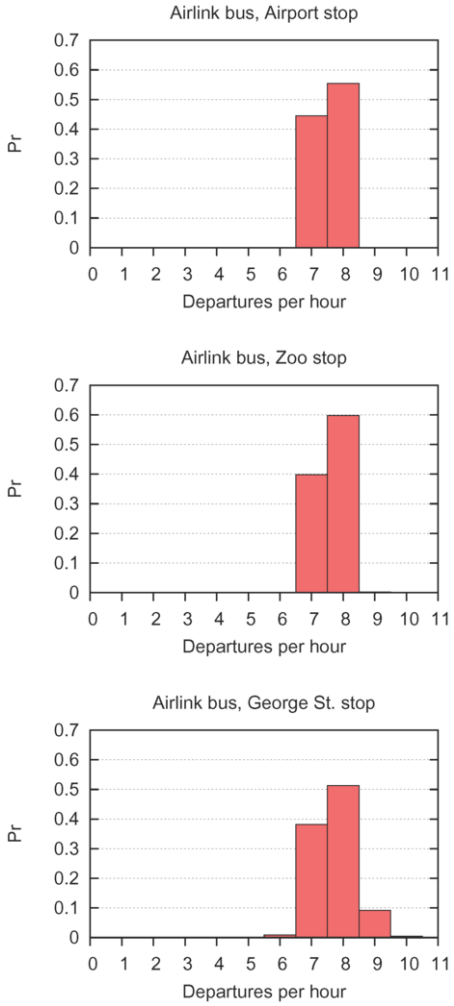

Fig. 21. Probability estimates of departures per hour for stops on the Airlink route, based on a single day simulated using the *runZ* algorithm.

Bus Tracker website. In [7], the AVL dataset of this paper was used to obtain bus sojourn time distributions of land patches in the Edinburgh city centre, which were then used to carry out 'what-if' analysis involving the introduction of trams. In [8], several statistical analysis techniques were used to evaluate the performance of several frequent services in Edinburgh (including the Airlink) in terms of the service level agreements discussed in Section 5, using the same AVL dataset.

8 Conclusions

In contrast to the results from a high-level model, measurement data has enormous authority. It is full of detail and quirks and seems to represent physical truth but as we have seen in examples above, it is not the whole truth, and it is not nothing but

the truth either. In our experience so far in working with data on the QUANTICOL project we have always needed to use human intelligence to clean the data before any automated processing could begin. An outlier only becomes an outlier when an interpretation is placed on the other data points.

What we saw in this example was that our understanding was enhanced by processing the data in a range of ways before any reflection and consequential adaptation took place. Further, there were complex collective percentile-based performance metrics to satisfy which required some ingenuity for us even to compute.

In some respects, our smart transport case study is relatively easy to work with. Data is readily available, and latitude and longitude data is relatively easy to interpret and visualise, allowing us to see problems in the data and apply data cleaning. We have intuitions about buses and transport, and local knowledge of what happens in practice. Further, we have access to the personnel in the Lothian Buses company who operate the system in practice. We can ask them what are the problems which are of concern to them. We have the potential to have some influence on the practice of the company, even if only a slight influence. Based on our calculations and more detailed reasoning [8], our belief at this point is that Lothian Buses are meeting the Traffic Commissioner's regulatory instruments.

In the future, we hope to use heat maps similar to the one in Figure 10 to automatically learn the bus routes, using a variant of the algorithm described in [9]. Using a representation of the routes in the form of a graph, we would be able to detect and remove outliers by removing measurements that are too far away from the edges on their route. This would allow us to automate the data filtering process, which at the moment is largely done manually.

Acknowledgements

This work is supported by the EU project *QUANTICOL: A Quantitative Approach to Management and Design of Collective and Adaptive Behaviours*, 600708. The authors thank Bill Johnston of Lothian Buses and Stuart Lowrie of the City of Edinburgh council for providing access to the data which was used for the case study. We would also like to thank Allan Clark for helpful comments on a draft version of this paper.

References

[1] Shao Yuan. Simulating Edinburgh buses. Master's thesis, The University of Edinburgh, 2013.

[2] Tim Pattinson. *Pruning GPS data with GPSprune*. Lulu, 2012.

[3] Smarter Scotland: Scottish Government. Bus Punctuality Improvement Partnerships (BPIP), March 2009.

[4] Margareta Friman. Implementing quality improvements in public transport. *Journal of Public Transportation*, 7(4), 2004.

[5] Peter Kemper and Carsten Tepper. Automated trace analysis of discrete-event system models. *IEEE Trans. Software Eng.*, 35(2):195–208, 2009.

[6] Ludovica Luisa Vissat, Allan Clark, and Stephen Gilmore. Finding optimal timetables for Edinburgh bus routes. In *Proceedings of the Seventh International Workshop on Practical Applications of Stochastic Modelling (PASM'14)*, 2014.

[7] Daniël Reijsbergen, Stephen Gilmore, and Jane Hillston. Patch-based modelling of city-centre bus movement with phase-type distributions. In *Proceedings of the Seventh International Workshop on Practical Applications of Stochastic Modelling (PASM'14)*, 2014.

[8] Daniël Reijsbergen and Stephen Gilmore. Formal punctuality analysis of frequent bus services using headway data. In András Horváth and Katinka Wolter, editors, *Computer Performance Engineering - 11th European Workshop, EPEW 2014, Florence, Italy, September 11-12, 2014. Proceedings*, volume 8721 of *Lecture Notes in Computer Science*, pages 164–178. Springer, 2014.

[9] Jonathan J. Davies, Alastair R. Beresford, and Andy Hopper. Scalable, distributed, real-time map generation. *Pervasive Computing, IEEE*, 5(4):47–54, 2006.